Accelerators in the Cloud

Users and resource providers perspective

Álvaro López García

<aloga@ifca.unican.es>

Advanced Computing and e-Science Group - Spanish National Research Council (CSIC)

May 7, 2019

Accelerated computing in the Cloud

- What do we mean with accelerated computing?
  - Generally: GPU access
  - Also: Other specialized devices such as FPGA, etc.
  - But also: Infiniband Interconnects
- Solved in the HTC infrastructure
- Still a pending issue in the Cloud

The users perspective

GPUs and Cloud users

- Not entering into community details
  - (For that see for example A. Bonvin plenary https://indico.egi.eu/indico/event/4431/session/3/contribution/45)
- What do users want? From our perspective
  - Increased usage of GPUs for Machine Learning, Deep Learning, Artificial Intelligence
  - GPU is requested as a first class resource (i.e. no GPU, no interest)
    - Not only at IaaS level, but also at PaaS
    - Users demand transparent access for cloud workloads (i.e. PaaS level)
  - Computation is done, sometimes GPU is not freed

The resource provider perspective

Accelerated computing in the Cloud

- Old story: Accelerators must be supported through the whole Cloud Stack
  - Hypervisor, CMF, Information Systems, Accounting, PaaS layer, etc.
- First and most problematic step: support at the hypervisor level

Hypervisor GPU support

- PCI-pasthrough works to an extent
- Most outstanding problems:
  - Need to fake CPUID instruction for (some) NVIDIA drivers to work
  - Whole device is being passed to the VM → Security concerns
    - See for example: https://medium.com/google-cloud/exploring-the-nuances-of-pci-and-pcie-7edf44acef94
- Peer-to-peer GPU virtualization in PCI-passthrough is problematic
  - PCI topology is lost, need workarounds to be implemented
- GPU virtualization not working in practice
  - Licensed products (that cost many €)
  - Limited combination of hypervisors and hardware

GPU sharing across workloads

- With PCI-passthrough the GPU is always occupied by the VM
  - Physical device attached to the VM
- GPU sharing within a VM is possible, but not very mature solution
  - E.g.: TensorFlow allocates a GPU memory, not allowing other processes to utilize it.
  - Cumbersome to set up memory quotas
- Leveraging NVIDIA MPS (NVIDIA CUDA Multi Process Service management program)
  - Different results, depending on the workload
  - Possible to run pure CUDA, problems with TensorFlow workloads
    - Need to specify the cpu_fraction to be used **beforehand**
- **Approach:** Share workloads at a higher level

Other technical drawbacks

- Stale VMs running after workload has completed
  - Resources not available, as VMs are consuming them!
  - Who should take care of this? VOs? EGI.eu? Resource providers?
- Suspension and migration are not possible
  - Physical PCI attached to a VM
  - Possible workaround: cold migration (stop, detach, migrate, attach, start)
- Proper driver must be configured through the whole set of tools (VMI, Docker container)!
- Security issues of PCI passthrough

Other not so technical drawbacks

- GPUs are only advertised in flavor's extra specifications
  - Cryptic properties like
- Still pending: Accounting
- Apart from GPU requests, limited interest in other accelerators
  - IFCA: Infiniband support since 2014, no requests coming from EGI.eu

Thanks for your attention

Any questions?

@IFCA_Computing